ClimateWins

Machine Learning Model Prediction of Weather Conditions & Climate Changes

By: ***Ang Wei Jie***

Date: ***August 2024***

Project Overview

Background

ClimateWins, a European nonprofit organization, is interested in using machine learning to help predict the consequences of climate change around Europe and, potentially, the world.

Objective

To investigate the various machine learning models available for the prediction of weather conditions using temperature data and identify the best machine learning model for the most accurate predictions.

Hypothesis

- Temperature data will allow us to prediction the weather conditions of the day.
- Supervised machine learning model will have high accuracy for prediction of weather conditions.
- Days with warmer temperature will lead to plesant weather conditions.

Data Sets

Data Source

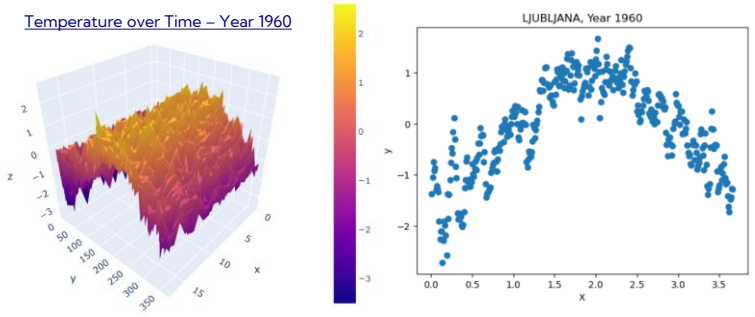
- Data collected by European Climate Assessment & Dataset (ECA&D).
- Contains weather observations data from 18 different weather stations across Europe from the late 1800s to 2022.
- Data variables includes daily value of cloud cover, wind speed, humidity, pressure, global radiation, precipitation, snow depth, sunshine, temperature.

Data Bias

- **Collection Bias:** The data was collect around 18 different weather stations across Europe, while, according to the European Climate Assessment and Dataset there are a total of 23755 weather stations across Europe.
- **Temporal Bias:** As the data ranges from 1800s to 2022, the data collected in the past might not be as accurate as the data collected in the recent years, due to the improvement in technology for data measurement tools.

Optimization of Data Sets

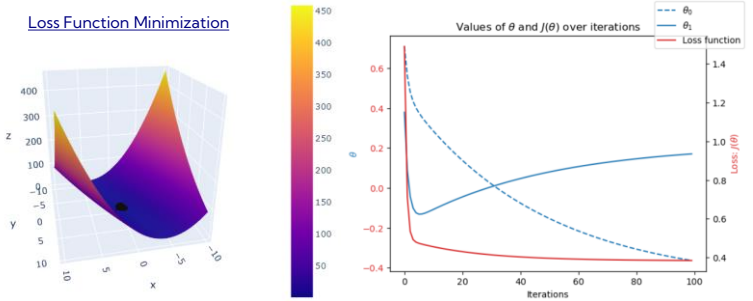
Temperature over Time – Year 1960



- Plotted data into a 3D visualization to view the temperature between the various weather stations throughout the year to get a general overview of the temperature data variation of the different weather stations.
- Extracted data from different weather stations of interests from specific years to further investigate the data points.

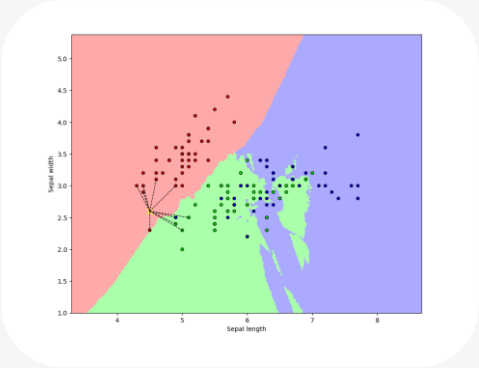
- Applied gradient descent to find the local minimum of the temperature data set.
- Optimize model performance towards finding the optimal solutions by minimizing loss function which assess the deviation between predicted and actual weather data.

Loss Function Minimization



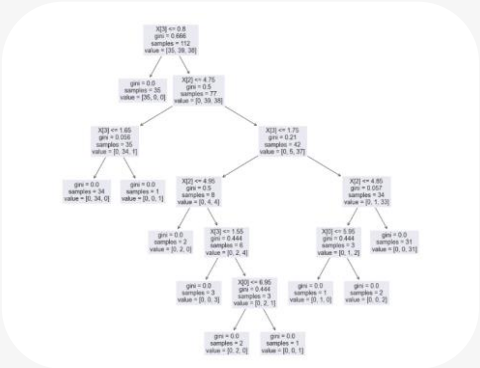
Supervised Machine Learning

K-Nearest Neighbours (KNN)



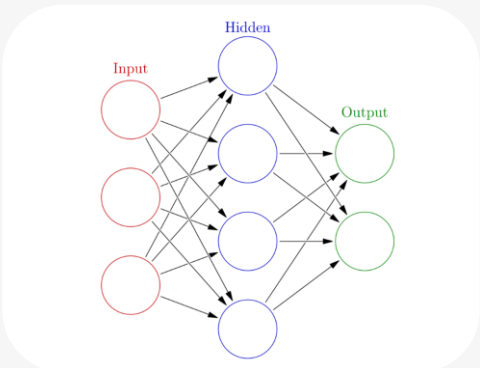
- Model uses a distance formula to determine which other point is the nearest, second nearest, third nearest, and so on (depending on the K value chosen).
- Model decides which group that data point is in based on how many of its neighbors belong to each category.

Decision Tree



- Model classify data with a list of forking branches that lead to an eventual answer.
- Each decision narrows down a solution by asking for more and more specific information.

Artificial Neural Network (ANN)



- Model takes a number of inputs and computes an answer based on a linear combination of all inputs multiplied by weights.
- Each input affects each node of the hidden layer, and each subsequent node affects all nodes on the next layer, until everything reaches the output layer.

Accuracy Comparison

K-Nearest Neighbours (KNN)

Weather Station	Accurate Prediction (Unpleasant Weather)	Accurate Prediction (Pleasant Weather)	False Positive	False Negative	Overall Accuracy	Unpleasant Weather Accuracy	Pleasant Weather Accuracy
Basel	3917	961	421	439	85%	90%	69%
Belgrade	3252	1544	524	418	84%	86%	79%
Budapest	3424	1462	476	376	85%	88%	80%
Debilt	4320	723	317	378	88%	93%	66%
Dusseldorf	4164	810	343	421	87%	92%	66%
Heathrow	4138	744	432	424	85%	91%	64%
Kassel	4563	614	252	309	90%	95%	67%
Ljubljana	3740	1180	455	363	86%	89%	76%
Maastricht	4253	824	309	352	88%	93%	70%
Madrid	2750	2261	418	309	87%	87%	88%
Munchenb	4237	792	309	400	88%	93%	66%
Oslo	4637	512	242	347	90%	95%	60%
Sonnblick	5738	0	0	0	100%	100%	N/A
Stockholm	4483	607	283	365	89%	94%	62%
Valentia	5404	74	58	202	95%	99%	27%
				Average	**88%**	**92%**	**67%**

- **Training Accuracy Score**: 0.5562
- **Testing Accuracy Score:** 0.4465
- **Overall Prediction Accuracy:** 88%

Parameter	
K Value	3

Decision Tree

Weather Station	Accurate Prediction (Unpleasant Weather)	Accurate Prediction (Pleasant Weather)	False Positive	False Negative	Overall Accuracy	Unpleasant Weather Accuracy	Pleasant Weather Accuracy
Basel	3871	943	467	457	84%	89%	67%
Belgrade	3183	1407	593	555	80%	84%	72%
Budapest	3394	1336	506	502	82%	87%	73%
Debilt	4289	749	348	352	88%	92%	68%
Dusseldorf	4106	832	401	399	86%	91%	68%
Heathrow	4092	759	478	409	85%	90%	65%
Kassel	4475	603	340	320	88%	93%	65%
Ljubljana	3679	1103	516	440	83%	88%	71%
Maastricht	4164	804	398	372	87%	91%	68%
Madrid	2828	2155	340	415	87%	89%	84%
Munchenb	4177	805	369	387	87%	92%	68%
Oslo	4542	549	337	310	89%	93%	64%
Sonnblick	5738	0	0	0	100%	100%	N/A
Stockholm	4422	601	344	371	88%	93%	62%
Valentia	5303	103	159	173	94%	97%	37%
				Average	**87%**	**91%**	**67%**

- **Training Accuracy Score**: 0.4617
- **Testing Accuracy Score:** 0.4726
- **Overall Prediction Accuracy:** 87%

Artificial Neural Network (ANN)

Weather Station	Accurate Prediction (Unpleasant Weather)	Accurate Prediction (Pleasant Weather)	False Positive	False Negative	Overall Accuracy	Unpleasant Weather Accuracy	Pleasant Weather Accuracy
Basel	4061	1051	277	349	89%	94%	75%
Belgrade	3363	1682	413	280	88%	89%	86%
Budapest	3568	1528	332	310	89%	91%	83%
Debilt	4418	803	219	298	91%	95%	73%
Dusseldorf	4177	935	330	296	89%	93%	76%
Heathrow	4272	872	298	296	90%	93%	75%
Kassel	4617	703	198	220	93%	96%	76%
Ljubljana	3790	1309	405	234	89%	90%	85%
Maastricht	4281	908	281	268	90%	94%	77%
Madrid	2882	2334	286	236	91%	91%	91%
Munchenb	4286	894	260	298	90%	94%	75%
Oslo	4725	566	154	293	92%	97%	66%
Sonnblick	5738	0	0	0	100%	100%	N/A
Stockholm	4463	822	303	150	92%	94%	85%
Valentia	5367	170	95	106	96%	98%	62%
				Average	**91%**	**94%**	**77%**

- **Training Accuracy Score**: 0.6325
- **Testing Accuracy Score:** 0.5019
- **Overall Prediction Accuracy:** 91%

Parameter	
Hidden Layer	100,200,20
Max Iterations	5000
Tolerance	0.0001

Summary

Insights

- **Artificial Neural Network (ANN) model** have the high prediction accuracy for prediction of weather conditions based on temperature data
  - ✓ Highest testing accuracy score of **0.5019.**
  - ✓ Highest overall accuracy of **91% correct prediction** across the different models.

- **Overfitting** of model **for Sonnblick weather station.**
  - ✓ All weather data points from this station are unpleasant weather conditions
  - ✓ Lead to model to **predict unpleasant weather condition 100% of the time**.

- **Lower prediction accuracy of pleasant weather** prediction may be due to the **lower amount of data for pleasant weather days.**

Recommendations

- To use would be the **Artificial Neural Network (ANN) model** for prediction of weather conditions based of temperature data.

- To **gather more pleasant weather data points** for specific weather station to enhance the prediction accuracy for pleasant weather day. The recommended stations are:
  - ✓ Sonnblick
  - ✓ Oslo
  - ✓ Valentia

- To **further optimize the parameter of the Artificial Neural Network (ANN) model** to be improve the prediction accuracy of the model.

Thanks!

Ang Wei Jie


weiji3x@gmail.com


linkedin.com/in/angweijie94

github.com/weiji3x